https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-518

DeViQ – A deep no reference video quality model

Steve Göring, Janto Skowronek, Alexander Raake; Dept. of Audio Visual Technology; Technische Universität Ilmenau, Germany; Email: [steve.goering, janto.skowronek, alexander.raake]@tu-ilmenau.de

Abstract

When enjoying video streaming services, users expect high video quality in various situations, including mobile phone connections with low bandwidths. Furthermore, the user's interest in consuming new large-size data content, such as high resolution/frame rate material or 360 degree videos, is gaining as well. To deal with such challenges, modern encoders adaptively reduce the size of the transmitted data. This in turn requires automated video quality monitoring solutions to ensure a sufficient quality of the material delivered.

We present a no-reference video quality model; a model that does not require the original reference material, which is convenient for application in the field. Our approach uses a pre-trained classification DNN in combination with hierarchical sub-image creation, some state-of-the-art features and a random forest model. Furthermore, the model can process UHD content and is trained on a large ground-truth data set, which is generated using a state-of-the-art full-reference model. The proposed model achieved a high quality prediction accuracy, comparable to a number of full-reference metrics. Thus our model is a proof-of-concept for a successful no-reference video quality estimation.

Introduction

Consuming video content over the internet has become such a success that most internet traffic is generated via streaming providers [7]. With todays availability of high-speed internet connections, users expect to obtain the best possible video quality, even in technically challenging scenarios such as low-bandwidth mobile phone connections or in rural heavily congested areas. To fulfill such expectations, different technologies such as new encoders or adaptive video streaming [16] are applied to keep the perceived video quality as high as possible. In the near future, however, the demand will further increase due to new content formats such as 4k/UHD resolution, high frame-rate, high dynamic range or 360 degree material.

To account for these challenges, an automated monitoring and optimization of the perceived video quality is an important factor for streaming services. For this purpose, different quality estimation models of different types (full-reference, reduced-reference, no-reference, and hybrid) have been developed. On the one hand, full-reference models are usually the most accurate ones as they compare the degraded material with the original reference material. On the other hand, no-reference models do not require the reference or – in case of a reduced-reference model – a parametrized representation of the original material, which makes these models particularly convenient for monitoring the service quality during normal operation. For that reason, this research is concerned with the development of a no-reference video quality model, with the goal to achieve a prediction performance comparable to state-of-the-art full-reference models.

The general idea is to build the model around a deep-neural-network (DNN). Current DNNs are able to outperform hand-crafted features or approaches for mostly every image or video related research question. We are using such a powerful pre-trained DNN in combination with some state-of-the art no-reference features to finally automatically create a model that is independent of resolution, future technology changes and does not rely on human annotated quality scores. We consider encoding artifacts because they are mostly relevant in adaptive video streaming.

This paper is organized as follows. In the next Section a brief overview of state of the art video quality models are described. In Section we discuss the main differentiating aspects of our model compared to the related work, while we describe our overall architecture in more detail Section . We conducted several experiments for evaluation of our proposed model in Section . Finally, we conclude with a discussion on the model and provide a short outlook and ideas for future work .

Related Work

There are many full-, no- or reduced reference video quality metrics reported in the literature. Some of them use DNN or machine learning techniques for computing video quality and can thus serve as basis for our model.

Torres Vega et al. [36] analyzed different video quality metrics under simulated network distortions, and found out that modern full-reference models are highly accurate to human perception. For example, Netflix's VMAF metric achieved quite good results compared to human ratings [23, 19]. VMAF is a compound video quality metric, it consists of several full-reference metrics (e.g. DLM [18], VIF [29]) and a per-frame motion estimator based on absolute average pixel difference. Using these generated values a Support Vector Machine (SVM) is trained to learn weights for calculating a combined quality score.

While VMAF achieved good results compared to human perception, a reference video is required, which is a drawback for non-intrusive quality monitoring of a service in normal operation. Looking at no-reference approaches as an alternative, Vega et al. described [37] a combined (based on machine learning) no-reference metric, which achieved a high correlation to the VQM full-reference metric. Thus, no-reference models using machine learning techniques can achieve comparable performance as full-reference models, an encouraging insight for our model.

Focussing on existing work using (deep) neural networks, a number of studies successfully applied neural networks for still image quality prediction [20]. Often, those approaches use patches of the input image to avoid large input layers for the neural networks. For instance, Kang et al. describes a no-reference image quality metric based on a convolutional neural network using patches of 32x32 pixels for images of 512x768 resolution [15]. Similarly, Dash, Mishra, and Wong also uses patches

(64x64 pixels) to train a DNN and in their experiments they achieved good classification results [8, 9]. Furthermore, there are other similar approaches for no-reference image quality estimation using patches in combination with DNNs or convolutional neural networks (CNNs) [17].

Next to no-reference approaches, DNNs are also able to perform quite good in quality assessment using reference material. For example, Bosse et al. [4, 3] used a DNN to calculate features from the reference and distorted images (again using patching) and to combine the extracted features to estimate the image quality. Alternative approaches to image quality prediction without patching have been reported in the literature as well [5, 14].

Thus, neural networks – with and without patching – can be used to predict image quality, both in a full- and no-reference manner. These insights are also encouraging for the quality estimation of videos, since many full-reference models that are not using neural networks are actually based on still image quality metrics (e.g. PSNR, SSIM, VQM, VMAF) [39, 23, 19].

Key aspects addressed with the model

Despite the encouraging results, we identified two open issues in the work just summarized and one further issue stemming from the basic properties of DNNs. As one can consider these as the main differentiators between our and state-of-the-art models, we briefly discuss them here before explaining the model in more detail in the next section.

A first issue concerns the aggregation of image quality scores across frames to a video quality score. In adaptive streaming applications quality switches can occur frequently, thus a simple averaging of each frame score is not suitable for all cases. While there is work on advanced temporal aggregation of video quality scores [28, 11, 31, 35], we will focus first on short term quality prediction. Thus we will consider video sequences with a duration of 10 s, a duration that is comparable with segments in a typical adaptive video streaming application [24].

A second issue concerns the patching of images, which is used to reduce the number of input layers of the neural networks. There are two aspects. First, using patching will divide each image in individual parts without any correlation or connection between them, meaning that global quality-relevant properties of the image can get lost. Examples are parts of a picture that span across different patches: a human would consider them as a whole when rating the quality, while an algorithm looking independently at individual image patches would not.

Second, patching is also not suitable for all cases from a computational perspective. In case of higher resolutions, processing time is quickly exploding as an increasing resolution requires an increasing number of patches to be included in the analysis. Consider the example of Kang et al. [15]. Using patches of 32x32 pixels on pictures with a 4k/UHD resolution of 3840x2160 pixels, such an approach would lead to 8100 patches for one single frame of a given video, compared to the 384 patches for the 512x768 pictures in [15]. We will address both aspects of patching by introducing a hierarchical patching approach.

A third issue concerns the advantage of DNNs in terms of dynamic feature extraction and the disadvantage of DNNs in terms of the large amount of required training data. Deep learning techniques allow to train robust models that do not depend on handcrafted features.

This allows to overcome one major disadvantage of handcrafted features, that is hand-crafted features are not able to scale for future technology changes without dedicated fine-tuning or extension. DNNs on the other hand can automatically extract features and can be re-trained dynamically.

However, training a DNN requires a large database and computational power. Thats why in some cases pre-trained DNNs are used for feature extraction. For our purpose, there exist several DNNs trained on huge databases that were successfully used for image classification or segmentation tasks. Such pre-trained DNNs are quite interesting for feature extraction, and we actually use one of them: the inception network described in [34].

Next to the feature extraction part, the pattern recognition part in our model requires a large amount of ground-truth quality ratings. For training the model, we will generate such ground-truth data from a state-of-the-art full-reference model: the VMAF metric [23, 19]. Note that our validation will be conducted against this ground-truth data in a cross evaluation fashion as well as against human quality ratings obtained a subjective quality assessment test.

Model architecture

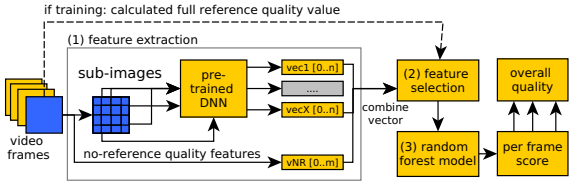

Figure 1. *General approach for training our model; using a pre-trained DNN plus some no-reference metrics to train a final model with scores from a full-reference video metric.*

The general model architecture is shown in Figure 1. The model consists of three steps: (1) feature extraction using a DNN with a hierarchical patching approach in combination with state-of-the-art no-reference features, (2) feature selection algorithm, and (3) quality score computation using a random forest model.

Step 1 uses a deep neural network (inception-v3 [34]) and two no-reference video quality metrics (BRISQUE [21] and NIQE [22]) for feature extraction. We use these two no-reference metrics to ensure that our model includes general quality related properties, both models are working in a similar pipeline, so that first features were extracted and later a regression model will be trained, we include only the calculated features. BRISQUE will calculate 36 distortion in-depend features (luminance based and using statistics for a spatial natural scenes) and NIQE one additional value for each frame as distance from naturalness.

Based on a given image, the pre-trained DNN will calculate a feature vector of probabilities. This vector is similar to what a user would perceive after looking at a given image from a cognition point of view. He or she would typically try to find and classify known pattern for an unknown image, each probability of the calculated vector is an indicator for such a known pattern/object. For each frame of a given video we calculate a feature vector v in the following way. We divide a frame in sub-images of equal size. The first sub-image is the complete frame, next are images of half of each dimension (4 images), followed by all images of 1/4 dimension (16 in sum) and last are all of 1/8 dimension (64 sub-images).

Summarized, for each frame we get a constant number of 64+16+4+1=85 images. These generated hierarchical sub-images can be seen as an iterative observation of a human. A human would first identify the general structure and later the fine details. The smallest sub-images has approximately the input dimension of our used DNN to ensure a correct classification. In contrast to state of the art patching approach this 85 hierarchical sub-image creation process is independent of input resolution. Patching or sub-image creation is needed, due to the fact, that current image classification networks re-scale images to a smaller resolution because of reducing complexity in calculation and the designed network. That's why a pure DNN for 4k/UHD resolution will increase computing time and needed memory in a drastic way.

Hence, we want to estimate video quality, such a re-scaling will infect quality properties. Compared to pure patching, we are able to enforce using our sub-image approach, that each sub-image is combined in another sub-image. Thus, a neighbor sub-images relation is modeled in another sub-image.

Each of these generated sub-images is then rescaled and applied to the DNN (we use inception-v3 network [34] for keras [6] with re-scaling to 299x299 pixels). A typical DNN classification network creates probabilities in the last layer for each origin object-class for that it was trained. In the case of the inception network 1000 class probabilities will be calculated, because it was trained for ImageNet Challenge which distinguish 1000 different objects.

$$pred(image) = \text{DNN probabilities of } image \quad (1)$$

$$no\_ref(frame) = \text{features from BRISQUE and NIQE} \quad (2)$$

$$dnn\_vec(frame) = [pred(\text{image}_0), ..., pred(\text{image}_{84})] \quad (3)$$

$$f\_vec(frame) = [dnn\_vec(frame), no\_ref(frame)] \quad (4)$$

We use the calculated class probabilities as features for our video quality metric calculation. For one frame, we generate in this setting a $85 \cdot 1000 + 37$ dimensional feature vector, compare Equation 1, 2, 3 and 4. We use two no-reference feature extractors ($no\_ref(frame)$) to extend the DNN feature vector ($dnn\_vec(frame)$) with 37 state of the art quality related feature values. This approach ensures that we do include some quality related calculations and not only perception based values.

In Step 2 and 3, we use the extracted feature vectors X ($feature\_vec$ for all frames of all training videos) with scores Y from a full-reference metric to train a machine learning model. We use as full-reference metric Netflix's VMAF implementation [23, 19]. Netflix's VMAF metric is widely used and itself a combination of several full-reference metrics.

Our machine learning model consists of a random forest regression approach with an additional feature selection step based on a extra tree regression. We tested several model parameters regarding feature selection (used threshold for identification of important features), number of trees or split criterion for the used random forest regression. As feature selection threshold we tested: *median*, *mean*, $0.5 \cdot mean$, $0.25 \cdot mean$ and $0.05 \cdot mean$, a threshold of $0.25 \cdot mean$ was the best. Furthermore, we tested the split criterion (*mse* – mean squared error, or *mae* – mean absolute error) where *mse* was the fastest and best performing. As last parameter we analyzed the number of used decision trees, due to the large training sample (approximately 100*k* frames), we tested $50, 100, 200, 400, 800$ and 1000 trees. We found out that 200 decision trees are the best trade-off between speed and accuracy, all other parameters for random forest regression and feature selection are default values from sci-kit learn [25]. Because of the large dimensional space, feature selection ensures that we only use important features in our resulting model. In the evaluation Section we will analyze the used features in detail.

Using the combination of a pre-trained deep neural network with other machine learning models, e.g. a random forest model, is an approach that is already used effectively in several other fields [12].

Training a random forest model can be done really fast in contrast to train a full DNN. Because of 4k/UHD resolution training a full DNN is difficult, so our hierarchical sub-image approach uses 85 images and creates a 85000 feature vector for each frame. Our prototype implementation is not optimized for speed, therefore we also need some time, comparable with the time that Netflix's VMAF calculation needs.

Evaluation

For evaluation of DeViQ we use several encoded video sequences for training and validation in a self created database. We are not using image quality assessment databases, e.g., LIVE II [30] or TID 2013 [26], because our general model should predict video quality for high resolutions, both datasets consists of images with low resolutions. Furthermore, there exists some video quality databases, e.g. Netflix Public Dataset [23] or VQEG HD3 Dataset [38], unfortunately they do not contain 4k videos, that is a requirement for our system. Because of the mentioned restrictions of public available datasets, we decided to create our own database, even we can extend our database easily.

First, we will describe our used video database, the training and validation sets. It consists of 360 distorted video sequences based on 12 different source video sequences with 4k/UHD resolution and mostly 60 frames per second. We use a 50% training and 50% validation approach based on source video sequences to ensure that our model gets completely new frames for validation. As second step, we compare the average overall quality scores with the corresponding VMAF values and other state of the art full-reference metrics. In our comparison we focus on full-reference metrics, because most no-reference metrics are not public available or not able to handle 4k video content. Further, we do not use pure image quality metrics, because our model is trained for video quality, mostly all state of the art image quality metrics are tuned and trained for databases with lower resolutions 4k and with different types of distortions. Our overall computation of video quality for a given video file is done using the same averaging approach that is used in VMAF-score calculation, this approach guarantees a comparison between both systems. In all of our evaluation experiments we consider various other full-reference metrics (PSNRHVS [10], MSSSIM [40], SSIM [41] and VIFP [29]). Additionally, we trained a random forest regression model without feature selection using 200 trees with only the BRISQUE+NIQE features, so that we can analyze the performance of our DNN features. We will answer the question how good our no-reference model can perform in contrast to a full-reference metric. Finally, we further conducted a subjective video quality test with 22 participants (average age=26.7) to estimate MOS values for each validation video sequence in an Absolute Category

Rating (ACR) approach [27] on the classical 5-point scale. Using these MOS values we are able, based on a linear mapping approach of our DeViQ-scores to MOS-scale, to evaluate our calculated overall mean scores with user's perceived video quality.

Dataset

Table 1: Our video database, T=for training, V=for validation

video sequence	source	T/V
MYANMAR	harmonic.com [13]	T
SINTEL_24FPS	blender.org [2]	T
SUGAR	TUIL	T
A_MYSTERIOUS_CASE	TUIL	T
CAMP	Sony [32]	T
MARKET_ELFUENTE	Netflix	T
AMERICAN_FOOTBALL	harmonic.com [13]	V
BIGBUCK_BUNNY	blender.org [1]	V
CUTTING_ORANGE	TUIL	V
VEGETABLES	TUIL	V
SURFING	Sony [33]	V
WATER_ELFUENTE	Netflix	V

As data source for our training and validation experiment we use high quality raw videos with 10 seconds duration, 4k resolution (3840x2160) and 60 frames per second (except for one video; SINTEL_24FPS 24 fps at 4096x1744). In Table 1 all selected 12 videos are presented, we will use 6 sequences for training (T) and the remaining 6 for validation (V). We use several sequences from various areas and sources to ensure content diversity. To sum up, we used two sequences from Blender (BIGBUCK_BUNNY, TEARS_OF_STEEL), two sequences from Harmonic (AMERICAN_FOOTBALL, MYANMAR), two from Netflix's El Fuente (WATER_ELFUENTE, MARKET_ELFUENTE), two from Sony's 4k-Demos (SURFING, CAMP) and 4 self created sequences. We encode each video to 30 distorted versions (3 different codecs; 4 resolutions, 2-3 bitrates per resolution), compare Table 2. In total, we created a dataset with 360 processed video sequences. Our overall dataset consists of approximatively 200k frames. For our training step we require quality scores based on a full-reference metric. That's why we also calculated VMAF scores for each frame of the 360 created video sequences. We split our training and validation dataset in a way, that each set has similar video sequences based on the shown content (animated sequences, much movement, less movement, static scene, changing scene, ...) and do not have similar frames. Furthermore in every of our experiments the validation video sequences were never seen by the system before, that means there is no overlap (also no frames with high content similarity) between training and validation videos.

Table 2: Encoding settings; in sum 30 settings per sequence

codecs	h264, h265, vp9			
resolutions	360p	720p	1080p	2160p
bitrates in Mbit/s	[0.2,0.75]	[1,2]	[2,7.5,15]	[7.5,15,40]

We analyzed in a small experiment our per-frame performance in comparison with VMAF scores, and found out that they have a similar *RMSE* performance to the overall video sequence performance. Therefore we will focus on per-sequence quality.

Per Sequence Overall Video Quality

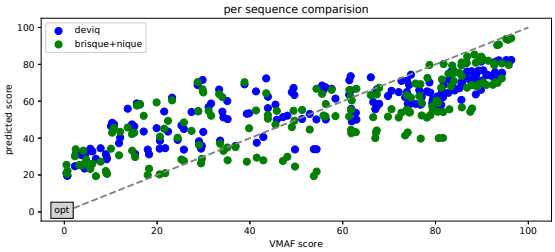

Figure 2. *Overall correlation of average VMAF-scores with* DEVIQ *and* BRISQUE+NIQE *predictions for each of the 6 video sequences.*

As a first experiment we will analyze the per sequence video quality. For taking better our general goal of prediction video quality into account, we applied the same approach as Netflix's VMAF for estimation of overall short sequence quality. We calculated the mean value of each frame quality score and archived quite good results. Figure 2 shows a scatter plot of our results for all validation sequences also in comparison with the used no-reference model. Both approaches, DEVIQ and BRISQUE+NIQE have a high correlation to VMAF scores.

Table 3: Comparison of our approach with different other full-reference metrics to the calculated VMAF scores.

method	*RMSE*	R^2	pearson	kendall	spearman
deviq	18.87	0.60	0.84	0.66	0.84
brisque+nique	19.75	0.56	0.85	0.64	0.83
vifp	22.28	0.44	0.58	0.46	0.63
msssim	48.99	-1.70	0.54	0.46	0.63
ssim	49.88	-1.80	0.48	0.44	0.60
psnrhvs	56.09	-2.55	0.33	0.52	0.72

For a more detailed analysis we further calculated *RMSE* (lower values are better), R^2 (coefficient of determination) (values > 0 show a linear correlation) and correlation coefficients: *pearson*, *kendall* and *spearman* (higher values show a better correlation). Our model performs quite well for these values, see Table 3. DeViQ outperforms other state-of-the art models, also the BRISQUE+NIQE baseline model, that we trained. However, we archive a higher correlation and less error in comparison with all analyzed full-reference metrics (i.e VIFP, MSSSIM, SSIM and PSNRHVS). DeViQ is approximately 15% better than the best full-reference metric (VIFP) considering *RMSE*. It also has a higher correlation to VMAF scores than every full-reference metric. The differences to BRISQUE+NIQE are quite small, however DEVIQ includes BRISQUE+NIQE features for calculation, that's why a similar performance is obviously. We just included BRISQUE+NIQE to analyze our sub-image feature creation approach, and it is notable, that our features improved the overall prediction accuracy.

Compared to other state-of-the art validation approaches, we are using a 50%-50% split of training and validation sequences. Thus our trained model performs well for unknown video sequences. Considering that our model is trained based on VMAF scores, a deeper analysis how accurate VMAF-scores and DeViQ-scores reflect MOS-values is required.

DeViQ and VMAF in Overall Video Quality with MOS values

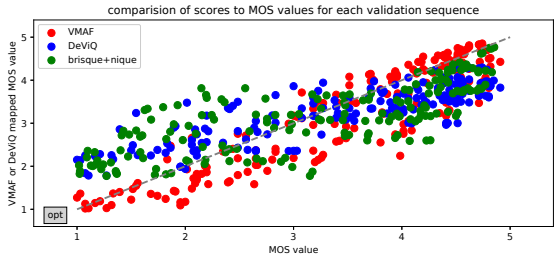

Figure 3. *Comparison of VMAF, DeViQ to MOS values, VMAF and DeViQ scores were mapped to a 1 to 5 scale.*

For each of our validation sequences, we calculated MOS values based on a conducted study for video quality perception. To map our estimated [0,100] scaled values to the corresponding [1,5] MOS scale, we applied a linear mapping function, $score\_to\_mos(score) = 1 + 4 \cdot score/100$. We analyze VMAF's, BRISQUE+NIQE and DeViQ's performance compared to these values (see Figure 3). All three systems show a high correlation to MOS values of our study. In a comparison VMAF scores show best matching.

Table 4: Statistical analysis of reference models and DeViQ to MOS values of our validation dataset.

method	*RMSE*	R^2	cohen_d	kendall	pearson	spearman
vmaf	0.55	0.76	0.24	0.72	0.92	0.89
deviq	0.70	0.61	0.19	0.61	0.84	0.81
brisque+nique	0.81	0.47	0.34	0.53	0.75	0.73
vifp	0.86	0.41	-0.34	0.52	0.70	0.67
msssim	1.70	-1.32	-1.72	0.46	0.69	0.61
ssim	1.74	-1.42	-1.76	0.45	0.65	0.60
psnrhvs	2.27	-3.15	0.30	0.60	0.34	0.76

To get a better overview we further performed a statistical analysis, considering *RMSE*, R^2, cohen_d, correlations (*kendall*, *pearson* and *spearman*). Table 4 summarizes all results. Best values for all analyzed statistics has the VMAF full-reference model, that's why we used it for generation of truth values in our training phase. DeViQ archives quite similar values than VMAF, however it does not rely on a reference video for calculation. Further, DeViQ is able to outperform all other full-reference models and on top of that also the BRISQUE+NIQE baseline model considering correlation and *RMSE*. Additionally, we calculated Cohen's d, DeViQ, VMAF and BRISQUE+NIQE have similar values that reflect a medium effect size.

Feature Importance

We further analyzed the feature importance based on which sub-images was used in our final model, see Figure 4. To get improve clarity we calculated normalized frequencies for each sub-image feature vector that reflect how often one of the feature values of this specific sub-image was important. In general our system uses 8.049 out of 85.037 feature values, so not all features are important. The top ten important features are spread over all created sub-images, that means our general hierarchically sub-image creation approach is useful for our final estimation. Most important features are for the smallest sub-images and middle ones. However, all importance values are quite similar, mostly between 5% to 25% of per sub-image group, the combination of them is required. Therefore our final model will use features of each of our created sub-images. Surprisingly, the most important feature-vector is for sub-image-28, that is one of the first images of our last layer in the hierarchically image-creation reflecting the different resolutions. E.g. for 4k/UHD resolution in comparison to HD these sub-images differ more in their quality, due to small granular details that are missing in the HD version.

Conclusion

Starting from the analyzes of the current state-of-the-art image and video quality models, we identified two main problems. First, most new quality measurement models (for images or videos) are using DNNs in combination with patching to reduce computing complexity. These patches are not able to handle global connections of a given image. Second, building up a self designed DNN for quality estimation can be difficult, because a huge database of human annotated per-frame data is required and the DNN-training is time consuming. We introduced a system called DeViQ (**De**ep **Vi**deo **Q**uality), that is able to handle both mentioned problems. It is based on a pre-trained DNN for feature extraction, state-of-the art no-reference features and uses a full-reference metric to automatically build up the training database. Therefore a large – human annotated – dataset for training is not required. Also, we combine the extracted features with a feature selection step and a random forest model, that can be trained quickly.

To tackle the patching problem, we use a hierarchical sub-image creation process, that ensures a almost global connection of each sub-image. In a large scaled evaluation experiment using a 50%-50% train-validation approach, we showed that our approach is able to perform better than various state-of-the-art full reference models. We archived high correlations compared to VMAF scores on per-sequence level with our no-reference model. Also the overall video quality prediction performance for our short sequences comparing with full-reference metrics is good. We found out that our model, that does not require any reference video, performs better for prediction of MOS values than full-reference models, except VMAF, however our system uses VMAF for training. Our prototype implementation is not optimized for speed, that's why further analysis in frame and sub-image selection should be conducted to reduce computing time, currently processing needs the similar time as VMAF calculation. Furthermore, we use a simple averaging approach for calculation of overall video sequence quality, there are more advanced approaches, e.g., considering movements, frame complexity, quality switches, possible. Moreover, using a wider database with more videos of different types would also increase accuracy and generalization of our model. We will analyze the open points in future experiments. To sum up, our system DeViQ can be used to train a no-reference model based on any full-reference model using any image DNN. We decided to use a classification DNN and VMAF full-reference scores and showed that DeViQ is able to outperform state-of-the-art full-reference models. Further, it is also possible to extend it to a full-reference model than can later be used for training DeViQ.

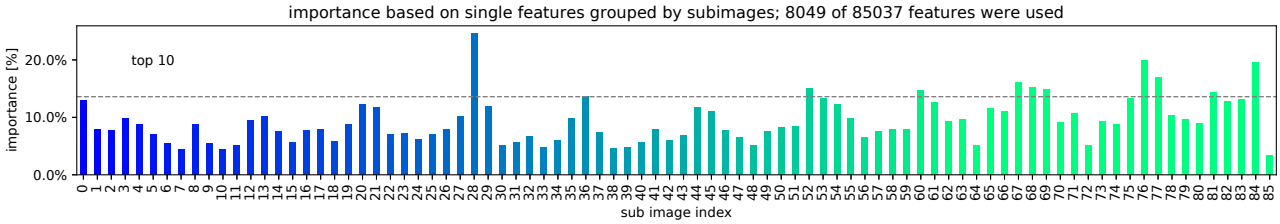

Figure 4. *Feature importance for each sub-images feature vector, image 85 represents the no-reference feature values.*

Acknowledgments

The authors would like to thank Michael Völske for running our experiments in a distributed fashion. This research work was partially funded by Telekom Innovation Laboratories, Deutsche Telekom AG, Germany.

References

[1] Blender Foundation. *Bick Buck Bunny Distribution*. URL: http://distribution.bbb3d.renderfarming.net/video/png (Accessed: 07/07/2017).

[2] Blender Foundation. *Sintel, the Durian Open Movie Project*. URL: https://media.xiph.org/sintel/sintel-4k-tiff16/ (Accessed: 07/07/2017).

[3] S. Bosse et al. "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment". In: *arXiv* (2016).

[4] S. Bosse et al. "Neural network-based full-reference image quality assessment". In: *Picture Coding Symposium (PCS), 2016*. IEEE. 2016, pp. 1–5.

[5] A. Bouzerdoum, A. Havstad, and A. Beghdadi. "Image quality assessment using a neural network approach". In: *Signal Processing and Information Technology, 2004. Proc. of the Fourth IEEE International Symposium on*. IEEE. 2004, pp. 330–333.

[6] F. Chollet et al. *Keras*. https://github.com/fchollet/keras. 2015.

[7] Cisco. *Whitepaper: Cisco Visual Networking Index:Forecast and Methodology, 2015-2020*. 2015.

[8] P. P. Dash, A. Mishra, and A. Wong. "Deep Quality: A Deep No-reference Quality Assessment System". In: *arXiv* (2016).

[9] P. P. Dash, A. Wong, and A. Mishra. "VeNICE: A very deep neural network approach to no-reference image assessment". In: *Industrial Technology (ICIT), 2017 IEEE International Conference on*. IEEE. 2017, pp. 1091–1096.

[10] K. Egiazarian et al. "New full-reference quality metrics based on HVS". In: *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*. Vol. 4. 2006.

[11] M. N. Garcia, W. Robitza, and A. Raake. "On the accuracy of short-term quality models for long-term quality prediction". In: *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE. 2015, pp. 1–6.

[12] R. Girshick et al. "Region-based convolutional networks for accurate object detection and segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 38.1 (2016), pp. 142–158.

[13] Harmonic. *Free 4K Demo Footage - Ultra HD Demo Footage*. URL: https://www.harmonicinc.com/4k-demo-footage-download/ (Accessed: 07/07/2017).

[14] W. Hou et al. "Blind image quality assessment via deep learning". In: *IEEE transactions on neural networks and learning systems* 26.6 (2015), pp. 1275–1286.

[15] L. Kang et al. "Convolutional neural networks for no-reference image quality assessment". In: *Proc. of the IEEE conf. on computer vision and pattern recognition*. 2014, pp. 1733–1740.

[16] D. K. Krishnappa, D. Bhat, and M. Zink. "DASHing YouTube: An analysis of using DASH in YouTube video service". In: *Local Computer Networks (LCN), 2013 IEEE 38th Conference on*. IEEE. 2013, pp. 407–415.

[17] J. Li et al. "No-reference image quality assessment using Prewitt magnitude based on convolutional neural networks". In: *Signal, Image and Video Processing* 10.4 (2016), pp. 609–616.

[18] S. Li et al. "Image quality assessment by separately evaluating detail losses and additive impairments". In: *IEEE Transactions on Multimedia* 13.5 (2011), pp. 935–949.

[19] J. Y. Lin et al. "A fusion-based video quality assessment (fvqa) index". In: *APSIPA, 2014 Asia-Pacific*. Dec. 2014, pp. 1–5.

[20] V. V. Lukin et al. "Combining full-reference image visual quality metrics by neural network." In: *Human Vision and Electronic Imaging*. 2015, 93940K.

[21] A. Mittal, A. K. Moorthy, and A. C. Bovik. "No-reference image quality assessment in the spatial domain". In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708.

[22] A. Mittal, R. Soundararajan, and A. C. Bovik. "Making a "completely blind" image quality analyzer". In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 209–212.

[23] Netflix. *Netflix VMAF*. URL: https://github.com/Netflix/vmaf (Accessed: 07/08/2017).

[24] R. Pantos. *HTTP Live Streaming*. 2011. URL: https://tools.ietf.org/html/draft-pantos-http-live-streaming-13 (Accessed: 07/07/2017).

[25] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *JMLR* 12 (2011), pp. 2825–2830.

[26] N. Ponomarenko et al. "Image database TID2013: Peculiarities, results and perspectives". In: *Signal Processing: Image Communication* 30 (Jan. 2015), pp. 57–77.

[27] I. Recommendation. "P. 910,Subjective video quality assessment methods for multimedia applications," in: *International Telecommunication Union, Tech. Rep* (2008).

[28] W. Robitza, M. N. Garcia, and A. Raake. "At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm". In: *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE. 2015, pp. 1–6.

[29] H. R. Sheikh and A. C. Bovik. "Image information and visual quality". In: *IEEE Transactions on image processing* 15.2 (2006), pp. 430–444.

[30] H. R. Sheikh et al. *LIVE image quality assessment database release 2 (2005)*. 2016.

[31] J. Søgaard et al. "Subjective analysis and objective characterization of adaptive bitrate videos". In: *Electronic Imaging* 2016.16 (2016), pp. 1–9.

[32] Sony. *Camping in Nature*. URL: http://4kmedia.org/sony-camping-in-nature-4k-demo/ (Accessed: 07/07/2017).

[33] Sony. *Surfing*. URL: http://4kmedia.org/sony-surfing-uhd-4k-demo/ (Accessed: 07/07/2017).

[34] C. Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* (2015).

[35] S. Tavakoli et al. "About subjective evaluation of adaptive video streaming." In: *Human Vision and Electronic Imaging*. 2015, p. 939407.

[36] M. Torres Vega et al. "An experimental survey of no-reference video quality assessment methods". In: *International Journal of Pervasive Computing and Communications* 12.1 (2016), pp. 66–86.

[37] M. T. Vega et al. "Predictive no-reference assessment of video quality". In: *Signal Processing: Image Communication* 52 (2017), pp. 20–32.

[38] VQEG. *HDTV Database*. URL: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx (Accessed: 07/07/2017).

[39] Y. Wang. *Survey of objective video quality measurements*. 2006.

[40] Z. Wang, E. P. Simoncelli, and A. C. Bovik. "Multiscale structural similarity for image quality assessment". In: *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. Vol. 2. IEEE. 2003, pp. 1398–1402.

[41] Z. Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.